DOCUMENT RESUME

ED 466 698 TM 034 277

AUTHOR	Vannoy, Martha
TITLE	Strategies for Identification and Detection of Outliers in Multiple Regression.
PUB DATE	2002-02-15
NOTE	15p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, February 14-16, 2002).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Heuristics; *Regression (Statistics)
IDENTIFIERS	*Outliers; Scattergrams

ABSTRACT

Outliers are frequently found in data sets and can cause problems for researchers if not addressed. Failure to identify and deal with outliers in an appropriate manner may lead researchers to report erroneous results. Using a multiple regression context, this paper examines some of the reasons for the presence of outliers and simple methods for identifying them. Heuristic data sets and scatterplots provide illustrations of the concepts discussed.(Contains 2 figures, 2 tables, and 11 references.) (Author/SLD)

Running head: OUTLIERS IN MULTIPLE REGRESSION

Strategies for Identification and Detection of Outliers in Multiple Regression

Martha Vannoy

University of North Texas

Paper presented at the annual meeting of the Southwest Educational Research Association, February 15, 2002, Austin, TX.

Abstract

Outliers are frequently found in data sets and can cause problems for researchers if not addressed. Failure to identify and deal with outliers in an appropriate manner may lead researchers to report erroneous results. Using a multiple regression context, the present paper examines some of the reasons for the presence of outliers and simple methods for identifying them. Heuristic data sets and scatterplots provide illustrations of the concepts discussed.

Strategies for Identification and Detection of Outliers in Multiple Regression

The presence of outliers can be problematic for researchers, therefore an understanding of the possible sources of these data points and strategies for detecting them are important for researchers to acquire. In 1999, Wilkinson and the American Psychological Association (APA) Task Force on Statistical Inference (TFSI) suggested, "The use of techniques to ensure that the reported results are not produced by anomalies in the data (e.g. outliers, points of high influence, nonrandom missing data, selection bias, attrition problems) should be a standard component of all analyses" (p.597). The report went on to say, "If you assess hypotheses without examining your data, you risk publishing nonsense" (p.597). Fox (1997) pointed out that unusual data can have a strong influence on regression analysis results. The TFSI noted that a single major value error could "make large correlations change sign and small correlations become large" (p.597). In light of these warnings, careful consideration should be given to identifying outliers and determining the proper course of action to take in dealing with them.

Therefore, the purpose of the present paper is to (a) discuss some potential reasons outliers may be present in a data set and (b) briefly illustrate some simple methods to identify outliers. To simplify the discussion, the paper uses multiple regression as the analytic context. Heuristic data is used to illustrate the concepts presented.

Reasons for Outliers

Outliers have been defined as data points that are "distinct or deviant from the rest of the data" (Pedhazur, 1997, p. 43), and "data points with unusually large residuals" (Evans, 1999, p. 213). They may also be seen as "cases that come from different populations than do most of the other cases in the sample" (Allison & Gorman, 1993, p. 156). Fox (1997) defined a regression outlier as "an observation with an unusual dependent-variable value given its combination of independent-variable values" (p. 276). It is important to note that, as Evans stated, "Data points that are outliers for some statistics (e.g., the mean) may not be outliers for other statistics (e.g., the correlation coefficient)" (p. 213). The data set in Table 1 illustrates this point.

INSERT TABLE 1 ABOUT HERE.

While case 5 is an outlier in terms of its effect on the mean and standard deviation, it falls on the regression line and does not function as an outlier in its effect on the correlation coefficient.

There are several possible sources of outliers. They may be due to the malfunctioning of an instrument, measurement errors, or recording or input errors (Pedhazur, 1997). The latter type includes mistakes such as misplaced decimal points or transposed numbers. Outliers may also be due to differences in conditions for the subject represented by the outlier (Gall et al., 1996) or a subject's conscious effort to sabotage a researcher's efforts (Huck, 2000). In addition, outliers may be a result of interaction effect where "individuals with a unique attribute or a unique combination of attributes reacting uniquely to a treatment" (Pedhazur, p. 44). Further, as Evans (1999) pointed out, outliers "may merely reflect natural variability within the population" (p. 214).

Detecting Outliers

With the acknowledgment that outliers may exist in a data set, the researcher takes on the responsibility of inspecting data for these extreme points. A complete discussion of outlier identification is given by Evans (1999). Therefore, only some of the more common and accessible methods are mentioned here for the applied researcher.

Researchers often equate extreme residuals with outliers (Pedhazur, 1997). Evans (1999) explained, "In regression analysis, a squared residual defines the amount of unexplained variability a given *i*th individual contributes to the total unexplained sum of squares, or the distance of the data point from the regression line on a scatter plot measured in the units of *Y*" (p. 216). Some researchers discard an observation if the size of its residual is greater than the estimated population standard deviation multiplied by a constant determined by the researcher. The larger the constant, the more likely the researcher is to retain observations with large residuals. Likewise, a smaller constant will result in the rejection of the largest residuals. The researcher must evaluate the risk involved in either retaining erroneous observations or rejecting valid ones (Evans, 1999).

Standardized residuals, or error scores converted to Z score form, are sometimes used as a means of detecting outliers because the standard scale facilitates interpretation (Evans, 1999). They are obtained by dividing each residual by the standard error of estimate. Pedhazur (1997) noted that some authors recommend scrutinizing standardized residuals with an absolute value greater than 2. Allison and Gorman (1993), however, suggested standardized residuals with an absolute value greater than 3 are possible outliers. In a normal distribution, the lower standard of +/- 2 could result in a larger number of potential outliers and present the researcher with the possible dilemma of rejecting valid observations.

Since the use of standardized residuals is based on the assumption that all residuals have the same variance (Pedhazur, 1997), the use of studentized deleted residuals may produce a more accurate picture of the outliers. Studentized deleted residuals measure the residual of the suspected outlier when its influence has been removed. A dependent variable outlier can be particularly influential, and if retained in the regression analysis, can pull the regression line towards itself.

Consider the data in Table 2 and the plots in Figure 1. With the 10th case included in the analysis using standardized residuals as in scatterplot (a), the regression line is pulled toward the outlier, thus indicating a relatively small residual. Noting the scale on the *Y* axis, when the analysis is calculated using studentized deleted residuals as in scatterplot (b), the distance of the outlier from the regression line indicates more clearly to the researcher the need to investigate the observation.

INSERT TABLE 2 AND FIGURE 1 ABOUT HERE.

Citing Hecht and Serdahl, Evans (1999) noted, "Outliers on the dependent variable typically exert greater influence on the parameter estimates and R^2 value than do outliers on the independent variables" (p. 220). Again citing Hecht, Evans went on to write, "Analyses of the standardized and studentized residuals were the most effective diagnostic methods for identifying outliers on the *Y* axis" (p. 220).

The scatterplots in Figure 1 illustrate the importance of using graphic methods to inspect data. Wilkinson and the TFSI (1999) noted, "Graphical inspection of data offers an excellent possibility for detecting serious compromise to data integrity" (p. 597). Scatterplots provide a reliable means of identifying outliers that are far from most of the data points (Evans, 1999). Fox (1997) maintained that numerical summaries can be misleading and that graphs are crucial to the effective analysis of data.

Another concept to consider in the detection and identification of outliers is leverage. The leverage statistic, also known as the hat value, identifies those cases that exert more influence than other cases on the regression model. Such high leverage points are outliers among the independent variables (Regression Analysis). Fox (1997) clearly illustrated the concept of leverage with three simple scatterplots, as seen in Figure 2. The outliers, represented by asterisks, have varying amounts of influence on the regression line. Because it is close to the mean of X, the outlier in the first graph (a) has low leverage and little influence. The outlier in the second graph (b), however, has high leverage and pulls the line toward itself. In the third plot (c), the unusual case has high leverage, but it does not alter the regression line because it is in line with the other data points. It would not be considered a regression outlier. A case that has a combination of high leverage and a large studentized residual has a strong influence on the regression coefficient (Fox).

INSERT FIGURE 2 ABOUT HERE.

The key point to be made by examining these strategies and methods for identifying outliers is that the process is important and should be undertaken. Millar and Hamilton (1999) described the process as separating data into a "subset of good ('reliable') observations and a subset of outlying observations" (p. 125), and the analysis is performed only on the “reliable” subset. The failure to include outlier identification in data analyses leads the researcher to risk basing findings on erroneous data. Of course, once extreme data points are identified, the researcher is faced with the decision of what to do about them. Removing outliers can have a serious impact on the regression model (Multiple Regression). Fox (1997) wrote, “Although problematic data should not be ignored, they also should not be deleted automatically and without reflection” (p. 285). Evans (1999) suggested, “To reject points simply because they are extreme is essentially to reject one of the assumptions upon which the regression analysis is based” (p. 226). Fox also said, “Truly bad data can often be corrected or, if correction is not possible, thrown away” (p. 285). Outliers can provide important insights (Huck, 2000) and may give the researcher cause to respecify the model and consider additional variables (Multiple Regression). Fox noted, however, that researchers need to “avoid ‘overfitting’ the data – permitting a small portion of the data to determine the form of the model” (p. 286).

When researchers take the time to examine their data for outliers and report the course of action that was taken upon discovering these extreme data points, the researchers should be commended for going the extra mile to ensure the integrity of their work.

References

Allison, D. B. & Gorman, B. S. (1993). Some of the most common questions asked of statistical consultants: Our favorite responses and recommended readings. *Genetic, Social & General Psychology Monographs, 119(2)*, 155-185.

Dallal, G. E. (2001). Regression diagnostics. Retrieved January 2, 2002, from Tufts University Web site: http://www.tufts.edu/~gdallal/diagnose.htm

Evans, V. P. (1999). Strategies for detecting outliers in regression analysis: An introductory primer. In B.Thompson (Ed.), *Advances in Social Science Methodology: Vol. 5* (pp. 213-232). New York: JAI Press.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Thousand Oaks, CA: Sage.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

Huck, S. W. (2000). *Reading statistics and research* (3rd ed.). New York: Addison Wesley Longman.

Millar, A. M. & Hamilton, D.C. (1999). Modern outlier detection methods and their effect on subsequent inference. *Journal of Statistical Computation and Simulation, 64(2)*, 125-150.

Multiple regression. (n.d.). Retrieved August 31, 2001, from http://www2.chass.ncsu.edu/garson/pa765/regress.htm

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth: Harcourt Brace.

Regression analysis. (n.d.). Retrieved January 2, 2002, from http://emlab.berkeley.edu/sst/regression.html

Wilkinson, L. & American Psychological Association Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

Table 1

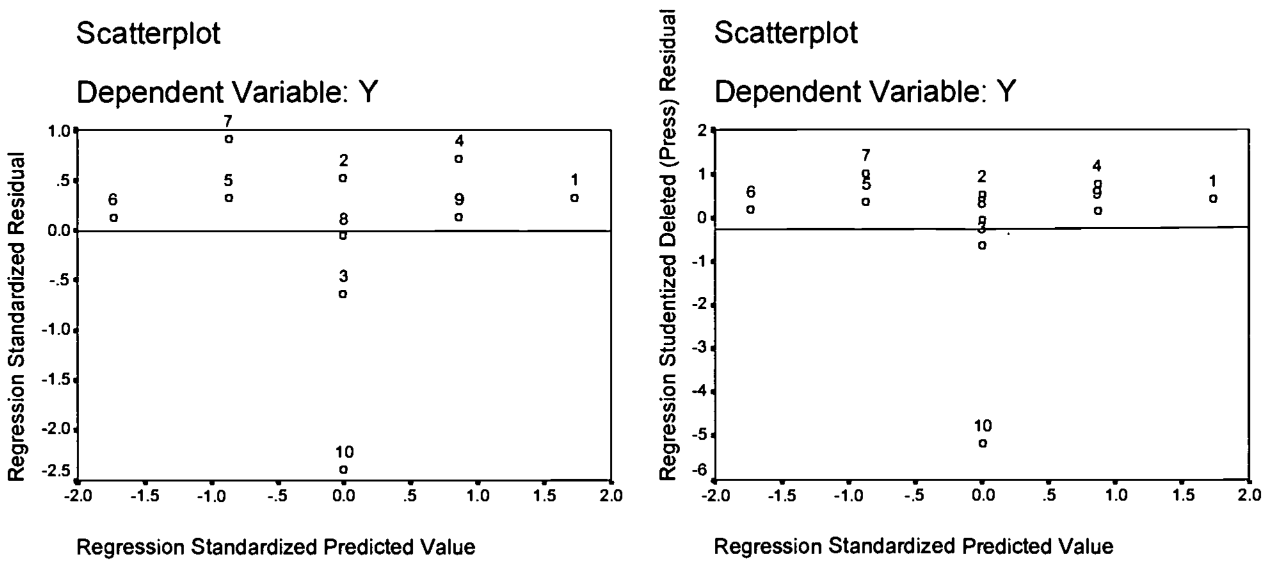
Data to Illustrate Impact of Outliers on Various Statistics

Case	X	Y
1	2	4
2	3	6
3	4	8
4	5	10
5	20	40
Mean	6.80	13.60
SD	7.46	14.93

Table 2

Data for Illustration of Outlier Effect.

Case	X	Y
1	10	10
2	8	9
3	8	7
4	9	10
5	7	8
6	6	7
7	7	9
8	8	8
9	9	9
10	8	4



(a)

(b)

Figure 1 A Comparison of Regression Analyses Using Standardized Residuals and Studentized Deleted Residuals.

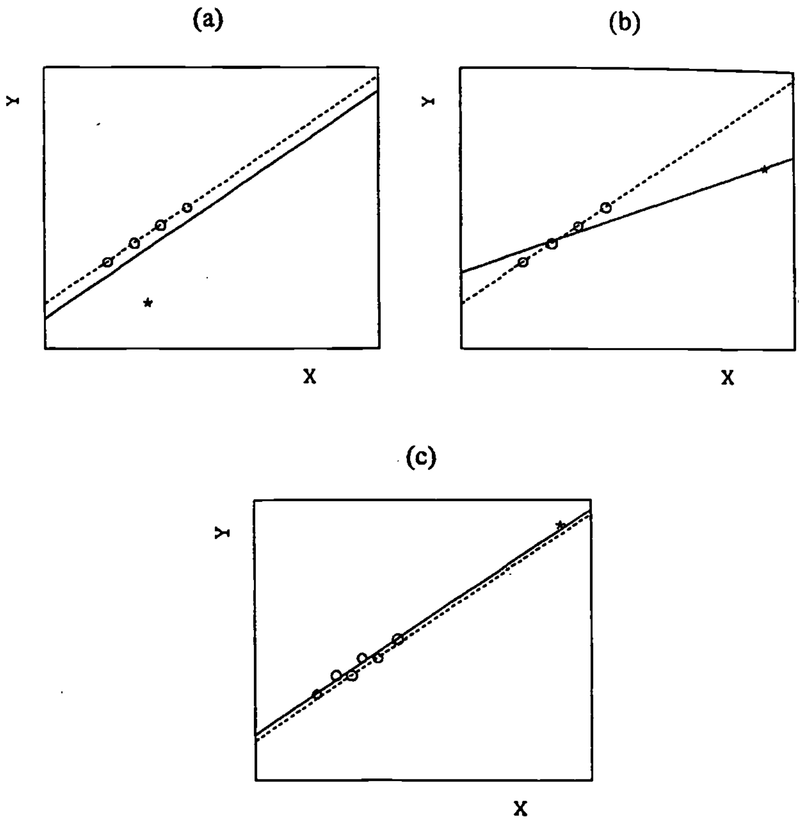


Figure 2 Examples of Outlier Influence in Regression Analysis.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

TM034277

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Strategies for Identification and Detection of Outliers in Multiple Regression	
Author(s): Martha Vannoy	
Corporate Source: University of North Texas	Publication Date: Feb. 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Martha Vannoy	Printed Name/Position/Title: Martha Vannoy / Res. Associate	
Organization/Address: 870 W. Buckingham Garland TX 75040	Telephone: 972-494-8514	FAX: 972-494-8951
	E-Mail Address: mvannoy@garlandisd.net	Date: 2-6-02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com